# Performance Analysis of IMS Signaling in Multimedia Networks

Nader F. Mir[*1], Sarhan M. Musa[2], Heng Gao[1], Chaitra Shivakumar[1]

[1]Department of Electrical Engineering, San Jose State University , San Jose, California, U.S.A

[2]Department of Engineering Technology, Prairie view A&M Univesity , Prairie View, Texas, U.S.A

[*1]nader.mir@sjsu.edu; [2]smmusa@pvamu.edu; [1]henggao@yahoo.com; [1]kschaitra84@gmail.com

***Abstract***

The deployment of IP Multimedia System (IMS) is on the rise and hence performance analysis og IMS becomes a critical area that needs to be investigated. One of the key areas to be addressed in this paper is the signaling part of the IMS architecture. IMS uses the Session Initiation Protocol (SIP) for its signaling during session establishment and termination. The nature of transactions involved in IMS is short and frequent hence signaling overhead needs to be analyzed. In this paper, we present the analysis of IMS signaling for some of the key SIP messages used in the IMS signaling. Also, the scalability and performance of the signaling portion are evaluated by using Network Simulator version 2 (NS2).

***Keywords***

*Multimedia Networks; IMS; SIP; Scalability*

## Introduction

IMS is defined by the 3rd Generation Partnership Project (3GPP) to enable the integration of cellular networks and the Internet. The 3G networks are fundamentally divided into two domains: the circuit switched domain and the packet switched domain. IMS is an overlay network over the packet switched domain of the 3G networks. It supports multimedia services including voice, video, audio and text transmissions.

IMS has emerged as the architectural framework for delivering multimedia services over IP. IMS was first proposed by 3GPP to support this capability of 3G networks. IMS adopts SIP architecture. As the IMS is based on SIP protocol for session control, it takes advantage of SIP features, such as being independent of access technology, and providing standard interface between the service plane and the control plane, etc. IMS provides the basis for service and network convergence [1].

The text-based SIP message is aimed to be easy to interpret and debug, but also adds a lot of overhead to the messages. When the number of User Equipments (UEs) increases, the use of text-based SIP messages for signaling in IMS could cause large number of signaling packets having to pass through in the network. This increase could lead to noticeable delay. Also, IMS was initially designed by 3GPP, which intended to support the multimedia services in mobile wireless network. The roaming function of wireless endpoints could increase the number of packets that need to be transmitted in the network. The roaming endpoints need to register with the servers in their home domain, and the following INVITE and other service requests also need to go through their home domain from the visited domain.

In addition, IMS multimedia services are provided by Application Servers. These Application Servers are not "pure" IMS entities, and can be distributed in different and possibly across geographically distant networks. Also the number of Call Session Control Functions (CSCF) servers to process requests in IMS network is much more than pure SIP network, and they are dynamically assigned to users. These characteristics could cause scalability issue as well.

Researches about IMS scalability issues have been conducted from different perspectives. One area of the study was focused on the CSCF configuration and message routing. In [2], the detailed functions of each CSCFs (P-CSCF, I-CSCF, S-CSCF) were discussed, and suggestions were given as to each CSCF to properly address the scalability issue. Another study put multimedia services into considerations while discussing the scalability issue. In [3], Presence Service (PS), one of the popular and fundamental services used in multimedia communication, was considered. After discussed the scalability issues with PS, the authors introduced partial solutions, including batched notification, common NOTIFY to multiple watchers, and optimizing federated presence with view sharing. Then they proposed their solution by

introducing a key component called inter-domain optimization module (IOM) to be deployed at each IMS domain so that it can interact with local PS to implement PS optimization together with differentiated QoS.

Both the studies are either theory-only or over IMS testbed deployment. Simulations of scalability issues in IMS were not covered, which could add new perspective to the discussion. This work uses NS2 to study the IMS scalability issue.

There are three main reasons why IMS is required: Quality of Service (QoS), Charging and Integration of different services. The packet switched domain in the 3G networks provides the real time multimedia services with best effort service and does not guarantee the QoS throughout a particular multimedia session. The network does not offer any guarantee on the bandwidth available for the session or the delay that the packets experience in the session. IMS takes care of the session establishment along with QoS provision so that users have a better experience while accessing real time applications [2].

The 3G standards bodies' 3GPP and 3GPP2 have both defined IMS with a few differences. IMS uses protocols defined by the Internet Engineering Task Force (IETF) to enable seamless working with Internet services. The 3GPP IMS makes it mandatory to use IPv6 whereas 3GPP2 IMS enables the use of both IPv4 and IPv6. IMS uses SIP for signaling and session management.

IMS provides a more reasonable method of charging for the services used. Currently, the charging is done on per byte basis. In this method of charging, the rates for real time applications can be really high as more number of bytes are transferred during each multimedia session. There is no way for the operator to determine the contents of each byte. With IMS this problem can be alleviated to a certain extent as the operator can charge based on the service used and not just on a per byte basis [2].

Indeed, the other significant advantage of IMS is reusability of third party services. Reusability of third party service makes it convenient to create new services by combining existing services without having to create them from scratch. This saves a lot of time for the operators and also reduces the time to market of the services. The aim of IMS is not only to provide new services but also to provide all the services that the Internet provides. IMS uses Internet technologies and Internet protocols to do this. The same protocol is used to establish a multimedia session between two IMS users, an IMS user and an Internet user and two Internet users [2].

In this paper, we mainly study the scalability of IMS with some analysis of signaling. The motivation for this study is an infrastructure used to support extensive and complex IP multimedia services in converged cellular and fixed networks with large number of users, IMS brings up scalability concern.

## Analysis of Scalability

Two areas regarding the IMS scalability are examined: first, the relationship between the number of UEs and the maximum delay; and second, the relationship between the bandwidth (both UE access bandwidth and bandwidth between servers) and the maximum delay. The following simplified IMS user registration flow is used for this study.

The registration delay here is defined as the time it takes from a user sending out the "REGISTER" request message to a P-CSCF, to the time the same user receiving the "200 (OK)" response message from the P-CSCF. As this performance study focuses on how the IMS core network will perform when there are lots of users trying to register at the same time, the maximum delay is used to measure this performance.

The maximum delay is defined as the longest registration delay among all the users that tried to register. Various numbers of users, different bandwidths (both of the access section and the core network section) were simulated to check the different outcomes.

The simulations are performed using Network Simulator version 2.34 (NS2), which is installed on an Ubuntu system release 10.04. The C++ code of NS2 SIP extension developed by Rui Prior [4] is modified to support the user registration flow stated in Fig. 1. All the CSCFs are implemented as SIP Proxys.

When receiving REGISTER request from UEs, P-CSCF server always forwards it to the I-CSCF; and it will send the 200 OK response message received from I-CSCF to the user who originates the registration.

For the I-CSCF, it will select the S-CSCF that serves the user where it is registered and forward the REGISTER message to that S-CSCF. As there is only one S-CSCF in each domain in the simulation, the selection is not implemented, and the IP address of S-CSCF is simply

set in code. Also, when S-CSCF sends back the 200 OK response message, I-CSCF will forward it back to the corresponding P-CSCF.

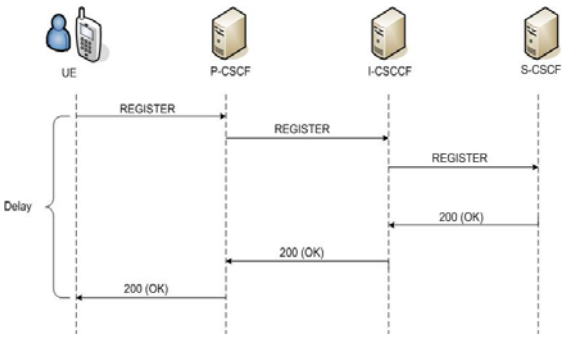


FIG. 1 PERFORMANCE SIMULATION SCENARIO

For S-CSCF, it will process the REGISTER request from I-CSCF. Assuming that all procedures are successful, the S-CSCF will then bind the user's physical address with its public SIP identity, and update the information in the HSS database. In the simulations, HSS is not implemented, and C++ list is used to create an entry for the user registration.

Tcl code is programmed to set up the network in the Fig. 2. In the Tcl code, two variables (number of UEs, and the bandwidth between servers) are set to receive from input. Thus, the same code can be reused for different simulation scenarios. Also in the simulations, it is assumed that network components have sufficiently large queue, in order to help prevent the possible drop of packets due to extensive delay.

The simulation network consists of two identical domains, each with the same number of UEs. When certain combinations of the number of UEs in each domain, the UE access bandwidth (from UE to the first router they get aggregated to connect to the core network) and the core network bandwidth (between server and the router interconnects them) are sent to Tcl from a Perl script file, Tcl script further sends these network settings and parameters to the ns shell to run the C++ code. First, the desired network topology will be built. And then UEs starts sending out REGISTER message to the P-CSCF, since P-CSCF is the single entry point in the system for all SIP messages from UE to IMS core.

After that, P-CSCF forwards the request to the I-CSCF as programmed. Once I-CSCF has received the REGISTER request, it forwards to the registrar - S-CSCF. After the S-CSCF registers the user, it will send back the 200 OK response using the same path as the request has come from, but in a reverse order, as shown in Fig. 1. Once the UE receives the response, time is measured according to the definition in the previous section, and the maximum delay is obtained for further analysis.

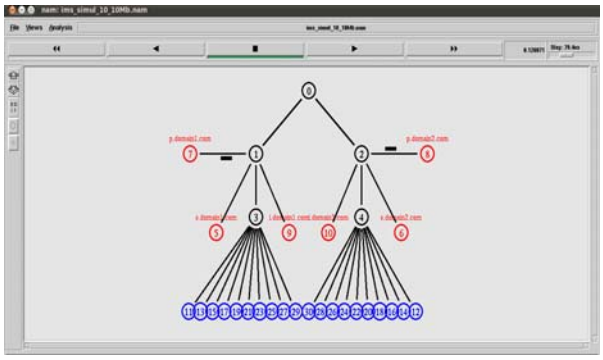

FIG. 2 SCREEN CAPTURE OF THE NS2 SIMULATION NAM OUTPUT

First, UE access bandwidth of 100Kbps is used to run the simulation. Core network (CSCF servers) bandwidth is set to 1Mbps, 10Mbps and 100Mbps respectively for each experiment. And the maximum delay is measured when changing the number of UEs from 10 and 50 up to 400, with step of 50. The text output trace file records the details of all packets passing through the network nodes. The results are shown in Fig. 3.

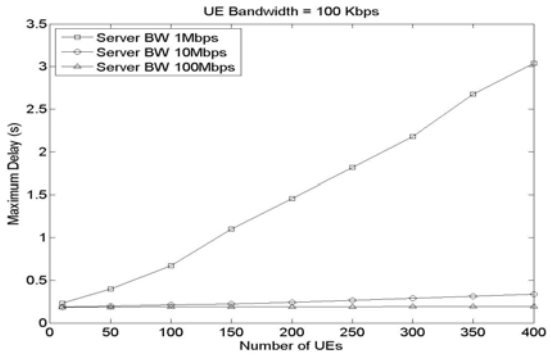


FIG. 3 MAXIMIUM DELAY VS. NUMBER OF UEs, UE BW = 100Kbps, 10 – 400 UEs

To better check the difference between 10Mbps bandwidth and 100Mbps bandwidth, further experiments are performed with UE numbers from 1 and 10 to 100, with step 10. The results are shown in Fig. 4.

Second, UE access bandwidth of 1Mbps is used to run the simulation. And core network (CSCF servers) bandwidth is set to 1Mb, 10Mb and 100Mb respectively for each experiment. And the maximum delay is measured when changing the number of UEs from 10 and 50 up to 400, with step of 50. The results are shown in Fig. 5.

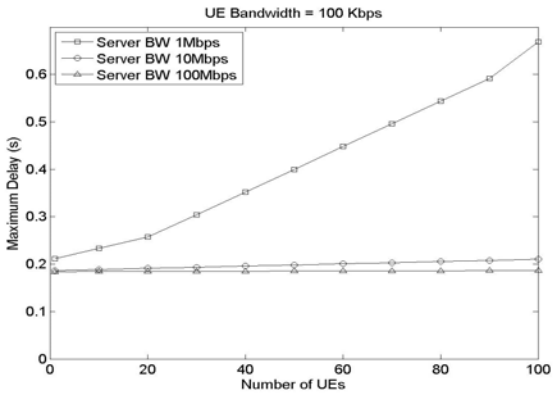


FIG. 4 MAXIMIUM DELAY VS. NUMBER OF UEs, UE BW = 100Kbps, 1 – 100 UEs

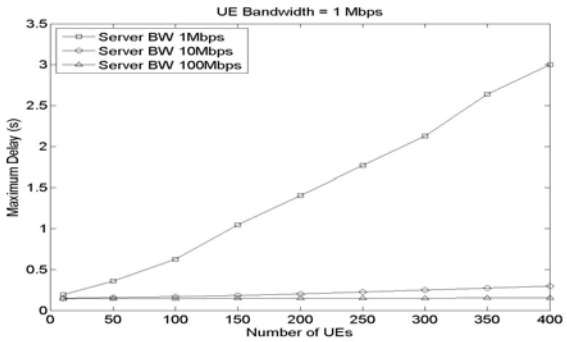


FIG. 5 MAXIMIUM DELAY VS. NUMBER OF UEs, UE BW = 1Mbps, 10 – 400 UEs

Similarly, to better check the difference between 10Mbps bandwidth and 100Mbps bandwidth, further experiments are performed with UE numbers from 1 and 10 to 100, with step 10. The results are shown in Fig. 6.

Based on the figures above, performance analysis is done regarding the scalability issue in IMS.

- Relationship Between the Number of UEs and Maximum Delay:

All the four figures above show that when the number of UEs increases, the maximum delay increases accordingly. It is obvious that when the network has more loads, and the network nodes (routers, CSCF servers) have increasing number of requests to process, so it will take longer time to finish all the processes.

- Relationship Between the Bandwidth and Maximum Delay:

Even though for all the bandwidth combination scenarios, the maximum delay increases as the number of UEs increases, the curves show different slopes or different shapes in different scenarios.

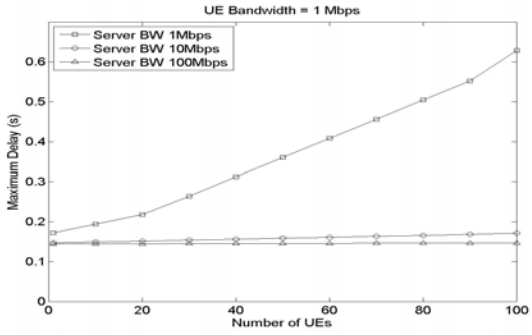


FIG. 6 MAXIMIUM DELAY VS. NUMBER OF UEs, UE BW = 1Mbps, 1 – 100 UEs

For the same UE access bandwidth, the maximum delay decreases when the core network bandwidth increases, and this decrease is more obvious when it changes from 1Mbps to 10Mbps, but less visible compared to the case when it changes from 10Mbps to 100Mbps.

For the same core network bandwidth, the maximum delay would increase when the UE access bandwidth reduces. But different UE access bandwidth has less obvious impact on the maximum delay. The difference is on a one hundredth second level.

## Analysis of Signaling Elements

The main functional element in the IMS architecture is the Call State Control Function (CSCF). A CSCF uses SIP to manage the multimedia sessions. There are three kinds of CSCFs: Proxy-CSCF (P-CSCF), Serving-CSCF (S-CSCF) and Interrogating-CSCF (I-CSCF). CSCFs are SIP servers [1]. P-CSCF - The Proxy-CSCF is the first point of contact in an IMS network. It accepts the SIP requests from IMS terminals or requests that are intended for an IMS terminal. It takes in the requests from the IMS terminal and either processes it or passes it on to another terminal [1].

S-CSCF – Serving-CSCF is a SIP server, which is always present in the subscriber's home network. The serving CSCF manages the session for the subscriber. It also acts as the SIP registrar and creates a binding between the user's SIP address and the user's IP address. All the SIP messages sent by or destined to IMS terminals go through the S-CSCF. The S-CSCF decides the action to be taken after deciphering the contents of the message. The S-CSCF can forward the SIP messages that it receives, to other SIP servers and CSCFs. The S-CSCFs can also interact with the Application Servers (AS) [1].

I-CSCF – Interrogating CSCF is the point of contact for external networks. It selects an S-CSCF for each

subscriber and routes incoming SIP messages for the subscriber to the selected S-CSCF. For example, the S-CSCF of a subscriber can be selected based on the capabilities requested, network capacities available, and network topological information. An I-CSCF can use encryption on the SIP messages to make the internal structure of the network transparent to the outside networks. This extra feature, which can be used optionally, is called "topology hiding internetwork gateway" [1].

SIP is one of the most popular signaling protocols for media over IP. SIP was originally designed in 1996, and in November 2000, SIP was accepted by 3GPP as a signaling protocol and permanent element of the IP Multimedia Subsystem (IMS) architecture for IP-based streaming multimedia services in cellular systems [5].

For IMS, the use of SIP signaling is not only between the IMS endpoints – UEs and IMS, but also among the IMS components – CSCFs, database, ASs.

SIP is designed to be a part of IETF multimedia data and control architecture [3]. It is an application layer signaling protocol for creating, modifying, and terminating sessions with one or more participants. These sessions include phone calls over the Internet, multimedia exchange, and multimedia conferences [4].

SIP has two network elements: Clients and Servers. Clients generate the SIP requests and the servers process the requests. There are several advantages to using the SIP protocol. Firstly, SIP is not concerned with the type of media that will be exchanged during the session. SIP is not restricted to a single transport protocol, but can be used over different transport protocols [3]. It is a text-based protocol and hence it is easier to debug. These designing features make SIP an ideal control/signaling protocol to facility the maximum interoperability between different networks, devices and protocols to carry various forms of multimedia sessions. The adoption of SIP as control protocol makes it possible for the IMS to be applied beyond the cellular networks which it initially intended for, and becomes an IP multimedia infrastructure that could bring the strength of both wired and wireless network together to build a converged framework for IP multimedia service.

Detailed SIP specifications are defined in IETF RFC 3261. In the RFC, the author explicitly states, "SIP does not provide services. Rather, SIP provides primitives that can be used to implement different services" [4]. This gives SIP the capability to support various types of multimedia applications/services, like instant messaging, presence service, push to talk over cellular services, etc.. There are extension RFCs define these services.

The signaling procedure in [1] represents the scenario where both the users UE1 and UE2 have roamed into a network other than their home networks. UE1, which is present in the visited, network 1 wish to connect to UE2, which is in the visited network 2.

It is assumed that both the users have registered with their respective home networks. The user UE1 begins the session by sending an INVITE message to the Proxy-CSCF in the visited network 1 (marked 1). The P-CSCF then sends the INVITE message to the S-CSCF that has been allocated for the user UE1 (marked 2 & 3). The S-CSCF then forwards the INVITE message to the appropriate I-CSCF (if the users are in different administrative domains) by checking the address part of the message (marked 4 & 5).

Once the I-CSCF gets the INVITE message, it queries the Home Subscriber Server for the location of the S-CSCF (marked 6) and sends the message to the S-CSCF (marked 7). The S-CSCF performs service control based on the UE2 profile. If the session is authorized, the S-CSCF forwards the SIP INVITE to the P-CSCF in Visited Network 2 (marked 8 & 9). The P-CSCF then sends the INVITE message to UE2. Once UE2 gets the INVITE message, it replies to UE1 with an Offer Response message.

The Offer Response message follows the same route as the INVITE message. UE1 determines the media streams offered depending on the information contained in the Offer Response. UE1 then sends an acknowledgement to UE2. Finally, UE2 can respond with an acknowledgment. Afterwards, UE1 then allocates resources to guarantee the Quality of Service for the session [1].

The signaling part of the IMS network was implemented using NS2. We integrated an open source SIP module into NS2. The P-CSCF, I-CSCF and S-CSCF are involved in the signaling. These are essentially SIP servers and hence the SIP module was extended for this functionality. The user session establishment and termination times were looked at for various values of bandwidth and delay. The sample topology involved users communicating through an originating P-CSCF, I-CSCF and S-CSCF and propagating through a terminating S-CSCF and I-CSCF on route to the receiver. The users in both

domains register with the SIP proxies in their respective domains.

DNS lookup/resolution was implemented and the users were placed in different domains using hierarchical addressing. The effects of increasing bandwidth and a huge delay on high traffic links were analyzed.

The very first thing that a user does is register with the proxy server. In this case it would be the P-CSCF. The P-CSCF is a stateless SIP server and does not keep a record of the user's profile or the state of the call. When the REGISTER message comes to the proxy, the proxy could either process the message itself or it could pass it on to another server.

Next, if the user wants to establish a session with another user in the same domain or a different domain, the user sends an INVITE message. Once the user establishes a session and exchanges data, the user terminates the session by sending the SIP message - BYE. Either user in a session can terminate the call. The SIP REGISTER, SIP INVITE and SIP BYE times were closely profiled across these varying bandwidth and delay links.

The implementation in NS2 was a simulation of the signaling portion of the IMS network. The only current simulator for IMS network is called Diabelli and the code is not open source. Future work would involve extending the simulation to the various stages of the voice call for both users in the home network and roaming users. Also various databases can be simulated for storing user information and session status. This would involve creating a new module for the other portions of the voice call and using a simple database.

User1 registers with the P-CSCF in its domain and the registration is forwarded to the I-CSCF and S-CSCF involved. User2 registers with the P-CSCF in its domain and the registration is forwarded to the I-CSCF and S-CSCF involved just as was done with User1. This completes the registration process. This was followed by a SIP Invite process, which again traverses through all the CSCF servers. The session was established between the two users. This completes the SIP session establishment phase. This was followed by a SIP termination phase. User2 started this and a SIP Bye message was sent. The time taken for session termination was tracked. Delay was kept constant and the bandwidth was varied to obtain the results below. Also bandwidth was kept constant and delay varied to get the second result.

As can be seen from the results, the effect of increase in bandwidth seems to saturate at some point. As seen in Fig. 7 as the bandwidth increases to around 25Mb the saturation point starts. This is explained by the fact that SIP signaling is more like a control plane message, which is only used for initial connection setup and teardown. The packet size of these messages is small in comparison with sending large amounts of data. It is also not a continuous stream unlike voice and video, which can show significant performance improvement with increased bandwidth.

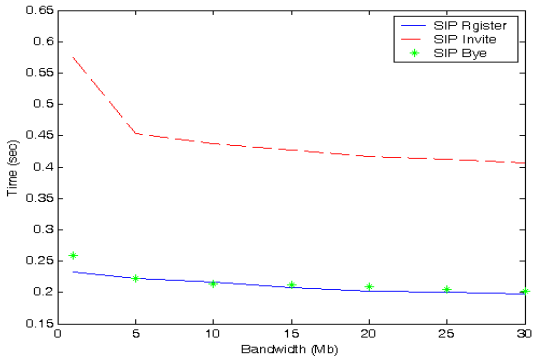


FIG. 7 SIP REGISTER TIME, SIP INVITE TIME, AND SIP BYE TIME VS. BANDWIDTH, WITH CONSTANT DELAY EQUAL TO 10ms

On the other hand we also see from Fig. 8 at about 25ms delay the time taken for REGISTER, INVITE and BYE messages increases significantly.

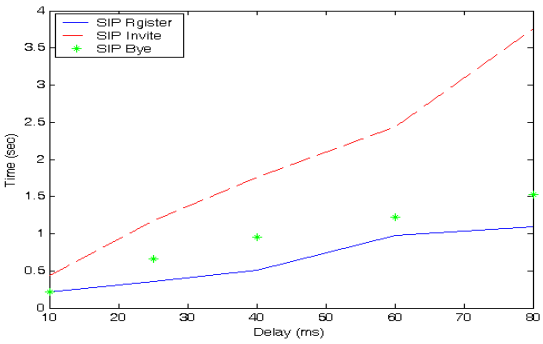


FIG. 8 SIP REGISTER TIME, SIP INVITE TIME, AND SIP BYE TIME VS. DELAY, WITH CONSTANT BANDWIDTH EQUAL TO 10Mb

Based on the above observation, we conclude that the IMS signaling phase using SIP is more prone to poor performance with large propagation delay. The bandwidth of around 25Mb is probably sufficient for a good performance for SIP signaling.

## Conclusions

The IMS signaling phases between the User agents and CSCF servers was simulated using NS2. The SIP signaling phases of REGISTER, INVITE and BYE were closely profiled across high-speed links and links with

large propagation delay. The signaling process is less tolerant to high propagation delays while large bandwidth has only limited impact on the performance of the signaling phases as seen from the results and the analysis.

## REFERENCES

Agrawal, Prathima et al., "IP Multimedia Subsystems in 3GPP and 3GPP2: Overview and Scalability Issues." IEEE Commun. Mag., vol. 46, no. 1, (Jan. 2008): 138–145.

Bellavista, P. et al.,. "IMS_Based Presence Service with Enhanced Scalability and Guaranteed QoS for Interdomain Enterprise Mobility." IEEE Wireless Communications, (June 2009): 16 – 23.

Camarillo, Gonzalo, and Garc´ıa-Mart´ın, Miguel A. "The 3G IP Multimedia Subsystem (IMS) Merging the Internet and the Cellular Worlds." John Wiley & Sons Ltd, 2006.

Collins, Daniel. "Carrier Grade Voice over IP", Second Edition, McGraw-Hill, 2003.

Issariyakul ,Teerawat and Hossain, Ekram. An introduction to network simulator NS2, New York; London: Springer, 2009.

IP Multimedia Subsystem http://en.wikipedia.org/wiki/IP_Multimedia_Subsystem.

Prior, Rui. "Universidade do Porto, ns-2 network simulator extensions," http://www.dcc.fc.up.pt/~rprior/ns/index-en.html.

RFC 3261., "SIP: Session Initiation Protocol." http://datatracker.ietf.org/doc/rfc3261.

Session Initiation Protocol, http://en.wikipedia.org/wiki/Session_Initiation_Protocol.

**Nader F. Mir** is currently a Professor, and served as the Associate Chairman, at the Electrical Engineering Department of San Jose State University, California. In the meantime, he serves as the Director of MSE Program in optical sensors and networks for Lockheed-Martin Space Systems Corporation for the University. He is also an expert in intellectual property development and a consultant for patent litigation cases in the areas of communications, telecommunications and computer networks at both the protocol and hardware levels. Dr. Mir has published two successful books, one of which is a world-wide adopted university text-book entitled "Computer & Communication Networks" published by Prentice Hall Publishing Co. He has also published numerous refereed technical journal and conference papers all in the field of communications and networking. He was granted a successful U.S. Patent in 2006, claiming an invention related to hardware/protocol for use in high-speed computer communication networks.

Dr. Mir has received a number of prestigious national and university awards and research grants. He is also the recipient of a university teaching recognition award and a research excellence award. He is also the recipient of a number of outstanding presentation awards from leading international conferences. He received the B.Sc. degree (with honors) in electrical engineering in 1985, and the M.Sc. and Ph.D. degrees both in electrical engineering from Washington University in St. Louis, MO, in 1990 and 1995 respectively.

**Sarhan M. Musa** earned his Ph.D. degree in Electrical Engineering from City University of New York, NY. Dr. Musa is currently an associate professor in the Engineering Technology Department of Prairie View A&M University, Texas. He has been director of Prairie View Networking Academy (PVNA), Texas since 2004. His research interests include computer communication networks and numerical modeling of electromagnetic systems. He currently serves on the Editorial Board of Journal of Modern Applied Science, and he is a senior member of the Institute of Electrical and Electronics Engineers (IEEE). He is also a Boeing Welliver and a LTD Sprint Fellow.